Automatic segmentation quality improvement for realization of unit selection speech synthesis

K. Szklanny †, M. Wójtowski †
† Multimedia Department, Polish-Japanese Institute of Information Technology, Koszykowa 86, 02-008 Warsaw

***Abstract* — This paper reports progress in the process of improving automatic segmentation. To help achieve natural sounding speech synthesis not only construction of rich database is important but also precise alignment must be conducted. Process of automatic segmentation often causes errors that needs to be corrected. Manual correction doesn't guarantee [1] that the process will be carried out consistently. Script for finding outliers was constructed and only necessary manual correction was prepared. Then a praat script was realized which allowed to detect and remove most important errors in automatic segmentation. Zero crossing process was applied to each phoneme. Additionally a comparison in prototype speech synthesizer was performed. The results shows that quality of generated speech has significantly increased.**

***Keywords* — automatic segmentation, unit selection speech Synthesis, Festival, improving alignment process**

I. Introduction

Manual segmentation of recordings, as a alignment process for the boundaries of recorded speech acoustic units is a very time consuming task, especially in the case of acoustic database designed for the purposes of speech synthesis, where the precision of the determined boundaries can decide about the synthesizer's success. Large amounts of speech signal that required to be divided into acoustic segments showed immediate need of automation of the process. Despite the fact that there are tools that cope with the task, their precision is far from sufficient as far as unit selection speech synthesis is concerned. Automatic alignment of the acoustical units boundaries is still an optimal solution as an introduction to further work on large speech databases [2]. Furthermore, the correction of the automatically determined segments can also be automated to some extent, thus allowing large amounts of speech signals to be processed within reasonable period of time.

There is a value in developing methods for high precision placement of phone boundaries [2].

The main aim of this study is to show the process of improving the quality of the automatic segmentation of a speech database designed for realization of unit selection speech synthesis. Creating of the database as well as generating of the automatic segmentation are described briefly below and in more details in [3]. The paper describes consecutive stages of segmentation corrections process as well as the testing in the prototype of the synthesizer.

This work is in part supported by the MNiSW under grant number N51600431/0641, author: MSc. Krzysztof Szklanny, kszklanny@pjwstk.edu.pl, Koszykowa 86, 02-008 Warsaw, tel.(0-22)58-44-562, fax (0-22)58-44-501

II. Automatic segmentation

The database was constructed [3] based on text corpus containing parliamentary statements [4] and newspapers reviews and rare polish words. The final corpus was balanced three times and the analysis shows that it is phonetically and prosodically rich [5]. It contains 2150 prompts.

The most common toolkits used for labeling acoustic databases are HTK and JRTk. These toolkits are based on Hidden Markov Models (HMM). Following topology was chosen: 3-state HMM models with 39 coefficients, namely 13 mel cepstral coefficients, energy and their first and second derivatives [6]. The polish database was automatically aligned with the transcription using an HMM based model trained with the HTK Toolkit. Different models were constructed. To choose the best structure of HMM test was performed based on 120 phrases. The final mixed Gaussian model was built on 585 phonetically balanced words, 25 ms analysis window with 10 ms frame period and estimated on automatically aligned database.

TABLE 1: Comparison between different structures of HMMs.

Model	Recognised words (%)	Recognised phrases (%)
1 ms frame period 5 ms analysis window	38,14	33,06
5 ms frame period 15 ms analysis window	71,47	55,65
10 ms frame period 25 ms analysis window	93,27	89,52
10 ms frame period 25 ms analysis window estimated on Unit Sel. Database	**92,95**	**89,52**
10 ms frame period 25 ms analysis window (diphone model)	71,79	53,23

Although the ratio of correct recognized words has fallen by 0.32%, as compared to the other models used, it was decided to use these models because the number of wrongly set boundaries in the database has fallen overall.

III. Error Detection and Manual Correction

A. Automatic error detection

The first stage of the automatic segmentation quality improvement was to automatically identify crucial errors and manually correct them. The scripts written in Perl and Praat [7] script languages were prepared in order to achieve it. Their tasks were: to check the duration of each phoneme in all the recordings, to count the global mean and the standard deviation, as well as to search and list occurrences of durations that differ from earlier estimated means at least by the doubled standard deviation (2SD) [8].

The scripts generated a list of about 4,500 abnormal in duration phonemes that required to be manually checked and corrected where necessary (about 1,400 different recordings). This simple technique made the identification of many crucial errors of segmentation and phonetic transcription possible.

B. Manual verification and correction

The manual correction of the segmentation was performed with Praat. The process concerned the recordings with phones listed as errors. First, the recordings were listened to and then verified the speech with transcription. Consequently, the manual corrections of segmentation as a whole were performed but mostly of the boundaries of listed phonemes were corrected.

In case of segmentation of a large speech database meant for unit selection speech synthesis it is crucial to be consequent in aligning the boundaries. This means that the boundaries of the same phonemes (or other acoustic units) ought to be set in the same manner. The consequence facilitates reaching the exact sound of the same segments of synthetic speech as well as its constant predictable quality. It should be noted that notorious errors in automatic alignment is a better scenario than variable segmentation because it is possible to predict how individual segments will be joined. Furthermore, even if it meant encountering the same error, it sometimes becomes possible to figure out a method of its automatic correction in all the recordings.

Characteristic errors that occurred repeatedly were identified at the beginning of the stage of manual corrections. The list of errors was prepared, which enabled limiting the range of manual corrections as well as studying the possibility of automatic correction of some errors was prepared. In order to achieve it almost 500 recordings were checked and identified errors were listed. A thorough manual correction of complete recordings picked by scripts would result in minor benefits as many recordings were excluded from the process. Because of possibility of no resemblance between the corrected segmentation and the one which was automatic, the quality of synthetic speech would depend on the used recordings. That is why it was decided to keep certain compatibility with automatic segmentation. In many cases it meant more liberal tolerance for the character of the set boundaries and sometimes even disregarding errors of minor importance. The mentioned tolerance mainly concerned most frequent errors that occurred on the prepared list. At this stage it was decided not to bother with placing boundaries at zero crossings. It's worth noticing that shortening the time needed for correction of one recording by a minute gave almost 24 hours of time savings (about 1,400 minutes).

Majority of the script listed errors concerned phonemes present at the ends of sentences or longer pauses in between the words. Parts of silence at the end of recordings as well as the pauses which were not identified were assigned to the neighbouring phonemes causing abnormal duration. Fig. 1 shows example of silence assigned to last phoneme in recording at top tier and corrected version at the bottom tier.

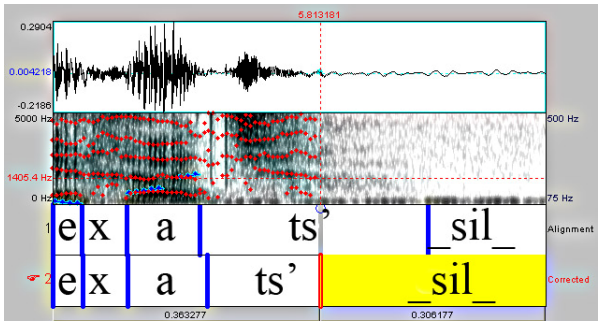

Fig. 1. Silence assigned to last phoneme.

It needs to be noted that those were not the most important errors as far as future use is concerned and in case of silence assigned to last phonemes they were of little importance. Although the scripts had successfully traced the most extreme errors they still partially disregarded the misalignments which were impossible to determine on basis of abnormal phoneme duration. However, thanks to the scripts it was possible to eliminate the most serious errors within reasonably short time with no need of thorough manual verification of all the recordings. Still, it seems that it would be more effective to use median instead of mean in order to search for segmentation errors. The reason being that distribution of phoneme duration is not normal distribution which means that most frequent values are not concentrated around mean values.

Fig 2,3 illustrates examples of manual corrections. Top tier is the automatic alignment and the bottom one with corrections included.

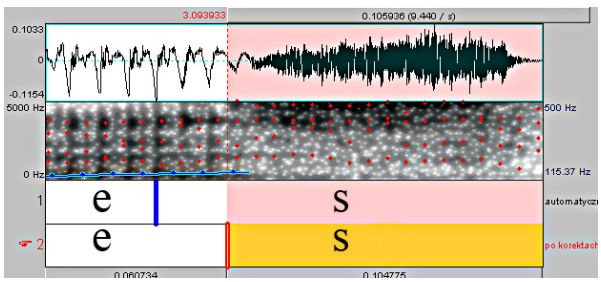

Fig. 2. Manual correction (phonemes /e s/).

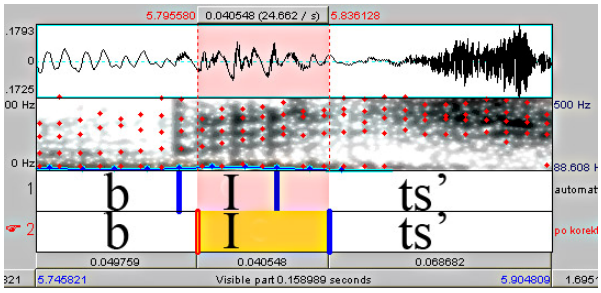

Fig. 3. Another manual correction (phonemes /b I ts’/).

C. *Frequent errors in automatic segmentation*

Identified frequent errors of the automatic aligner were as follows:

- Voiced plosive phonemes /b/, /d/, /g/ + vowel were too short, the voiced part of plosive phoneme was assigned to vowel
- Voiceless plosive phonemes /p/, /t/, /k/ usually started in the second part of the previous phoneme, and often finished too early (sometimes it was only part of previous phoneme + silence)
- Voiceless plosive phoneme followed by fricative (ie. /pS/, /t S/, /ks’/, /ps/) - sometimes a part of fricative was assigned to plosive phoneme
- vowel + fricative: /s/, /S/, /Z/, /z/, /s’/ - part of vowel was assigned to following consonant
- Affricate phonemes /ts’/, /tS/, /ts/, also in the neighborhood of vowel – in many cases starting boundary of affricative was placed before ending of preceding phoneme – if affricative was followed by vowel, finishing boundary of consonant was often set in the final sound of vowel
- /l/, /w/, /r/, and /v/ - too short segments
- The same phonemes one after another (geminates) ie. /rannI/ - If they had not been clearly separated by the speaker, they were almost entirely marked as one phoneme hence the other one contains merely a small remaining part.
- Last phone in the recording – a part of silence was joined with the previous phoneme
- Long silence between words – the whole silence was joined to neighboring phonemes [9]

IV. Issues

A. *Zero crossing*

Because of the usage of speech database for realisation of unit selection speech synthesis in Festival [10] the boundaries of all the phonemes should be aligned at signal zero crossing. The precise reason is the lack of modification of the joined segments (multisyn voice). Otherwise it could cause discontinuities of amplitude in synthetic speech and related crackings. The condition was not fulfilled by automatic segmentation as the used HMM models had a 10ms frame period, which is too long to find zero crossings. Unfortunately the presented above statistics showed that shorter frames did not enable reaching correct phoneme boundaries. That is why it was important to prepare a Praat script which automatically moved the boundaries of all the phonemes to the nearest possible amplitude zero crossing. Furthermore it was decided that the process will be performed after manual correction of the recordings traced by the scripts. The advantage of such work schedule was the possibility of skipping the time consuming task of placing boundaries in zero crossing during manual corrections as well as studying the possibilities of introducing automatic corrections on basis of automatic segmentation errors observation. Thus, also the recordings that did not undergo manual correction could be dealt with and the manual correction could be limited to some extent.

B. *Distortions*

Significant distortions in the range of 50hz frequency related to the operating of AC were identified in some recordings during manual verification of segmentation. The distortions occurred as a variable shift of signal in relation to zero, reducing the possibilities of zero crossing in some parts. In some cases those could also be distortions coming from the surroundings and caused by devices using fanlike components such as air conditioning, hand dryers or ventilation. The studio where the material was recorded is located within the institute grounds and the recordings were conducted during the academic season.

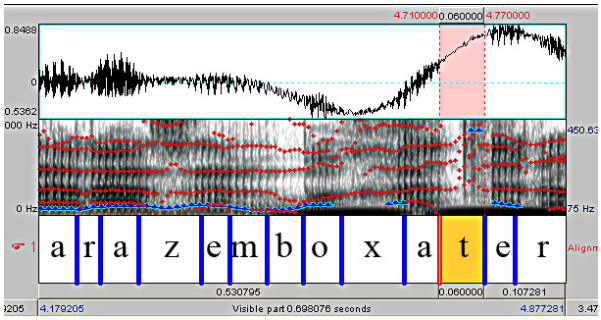

Fig. 4. Example of 50 Hz distortions.

Sound files with identified distortions which could affect the quality of speech synthesis was listed during manual corrections. It should be noted that manual verification and correction concerned only traced by scripts recordings with abnormal in duration phonemes. That is why it was decided that recordings excluded from the manual check needed to be verified automatically by correction script described below.

V. Correction Script

There is a need for improving the process of forced alignment. One method is presented in [11]. It is based on creating regular expression module for post-lexical rules of phonetic transcription. Also in IVONA polish speech synthesis system automatically labeled database was additionally processed to resolve pause disambiguation [12]. It was decided to develop praat script in order to move all the phonemes boundaries to zero crossings and to check recordings excluded from the manual corrections if they contain AC distortions for improving the quality of the alignment. Additionally it was an occasion to examine

possibility of introducing automatic corrections of chosen misalignments of segmentation. Moreover, eliminating the need of manual verification of all recordings after automatic correct process was of key importance. That is why correction script included mechanisms which made verification of every introduced automatic change and reporting of uncertain cases possible.

The process of creating the script was divided into two stages. First algorithm for moving boundaries to zero crossings and verification procedure was prepared. This stage also included testing process. Verification criteria for automatic corrections was the distance between new phoneme boundary set by the script and its previous position. If the distance was bigger than specified safe range given phoneme was listed, in extreme situations previous boundary positions was additionally kept unchanged. It needs to be noted that verification procedure based on the phoneme boundaries shift coped also with identification of significant AC distortions that caused local shifts of signal in relation to zero, as well as lack of zero crossings in some parts. Such distortions in most cases forced the script to try to set actual phoneme boundary even further than boundaries of the following phoneme which is far further than the permissible range and, in consequence, the recording was listed as needing manual verification.

The second stage of creating the script was to develop mechanisms that made some other automatic corrections as well as their verification possible. This stage was preceded by studies of introducing automatic corrects and verification of chosen listed frequent automatic segmentation errors. On the basis of investigation followed by tests the decision was made to automatically correct only voiceless plosive phonemes (/p/, /t/, /k/), whose starting boundaries were almost always set in the second part of the preceding phoneme and also normally ending boundaries were set to early (without plosive part). Those are some examples of the most important errors on the prepared list because of occurring cases where phoneme boundaries contained only a part of the previous phoneme along with silence which were unacceptable as far as speech synthesis is concerned. Additional reason was relative easiness of describing, implementing and verification of voiceless plosive phonemes corrections. It should be noted that voiceless plosive phonemes are preceded by a very significant fall of energy (silence) just before their start, which is followed by significant impulse of energy just before their end. This fact enabled the possibility of conducting corrections in a relatively easy and safe way, on the basis of tracking energy changes. Despite the fact that at this stage only 3 phonemes were corrected, a time-consuming test was needed in order to find proper parameters values that minimized the risk of undesirable modifications but also guaranteed correction of important errors. The problem was in the large number and differentiation of the recordings. It means multitude of contexts in which phonemes were used, their different duration and different energy level.

The script generated a list of over 700 phonemes that needed to be manually verified and corrected where necessary. During the verification of listed phonemes segmentation as a whole was also briefly checked.

Fig 5 shows an example of a phoneme boundary moved by the correction script to the nearest positive zero crossing. Fig 6 illustrates an example of automatic correction of phoneme /t/ carried out by the script.

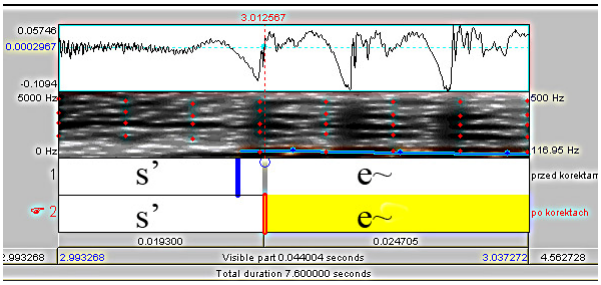

Fig. 5. Example of phoneme border automatically moved to positive zero crossing.

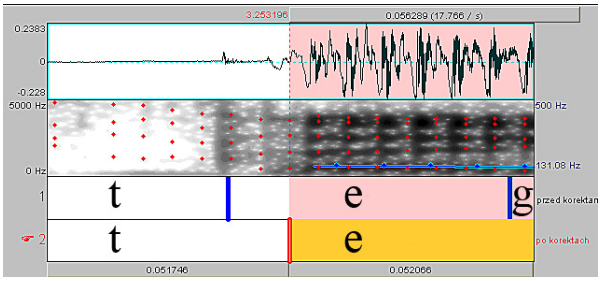

Fig. 6. Automatic correction of plosive phoneme /t/.

VI. DISTORTIONS REMOVAL

AC distortions described above were removed with high-pass filter using open source Audacity sound editor. This process concerned recordings listed during manual corrections and by correction script. Fig. 7 shows Audacity's window with distorted signal at the top and the same signal after filtration below.

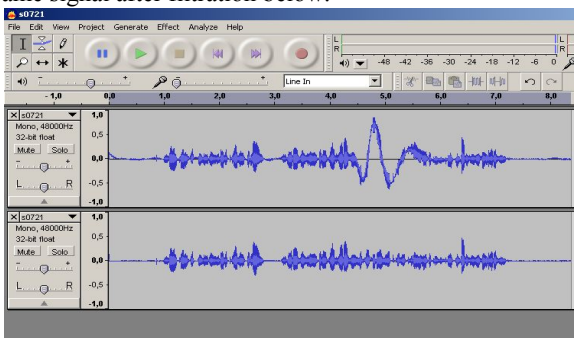

Fig. 7. Distortions removal.

VII. TESTS

Testing and the verification procedure of correctness of the segmentation were the last stage of the present work. It was the chief objective to estimate the quality of the database before and after the introduced amendments. The tests were conducted in the prototype of Polish unit-selection synthesizer. The quality of speech generated by the synthesizer was the determinant of correctness. The

main factor was natural sounding of joined acoustic units, especially in the joined places. It was searched for incorrect segments, clicks and other distortions, which could be caused by wrong segmentation.

A. *Testing procedure*

The procedure of testing and the verification included generating and listening to 100 sentences from the test corpus discussed below but also significant number of other randomly chosen phrases and sentences.

About 30 sentences was generated using both the data before and after their correction. Generated sentences were listened to by two phonetic experts. The main purpose was to find out if the implemented corrections had improved the quality of segmentation and also the quality of the generated speech [13].

B. *Corpus for tests*

The purpose of creating the corpus was to obtain a set of sentences which will meet specific requirements different from the ones that were used to create the main corpus [14]. It was decided to obtain a very small corpus and at the same time the biggest possible coverage of different acoustic units and their connections as well as selection of sentences of different subjects from the ones included in the acoustic database. The variety of corpora was supposed to ensure testing the naturalness and comprehensibility of generated phrases occurring occasionally in the main corpus.

The test corpus was prepared in the CorpusCrt application. Sentences was compiled from three different linguistic bases, containing texts from newspapers of various subjects. Before the test corpus was created, it was required to generate the phonetic transcription for phonemes diphones and triphones for the whole database. The phonetic transcription was generated with the Perl scripts. It was decided to limit the size of the test corpus to 100 short statements (max. 60 phonemes in each sentence). The reason why the size of the corpus was kept to the minimum was that, apart from the corpus itself, there was a large number of sentences and expressions tested at random.

The criteria of the sentence selection referred to their maximum length, the quantity of occurrences of various acoustic units and different phoneme configurations. During corpus balancing it was decided that each phoneme should occur at least 25 times, each diphone and triphone should occur at least once . Because of the small size of the corpus, obtaining all the diphones and triphones was impossible, however, required condition ensured variety of occurrences of mentioned acoustical units.

C. *Results*

The improvement of the segmentation process was verified on the basis of perceptual test and comparisons of sentences generated with automatic and corrected segmentation. At this stage the most objective was not the quality of the generated speech but obtaining the sentences with the minimum number of segmentation errors. The test showed that quality of generated speech has increased and the distortions number was significantly decreased. In 90% of cases the difference was to be heard clearly. Fig. 8 and 9 shows the difference in generated speech. Sentence ‘Sprzed domu twierdząc’ was synthesized unintelligible.

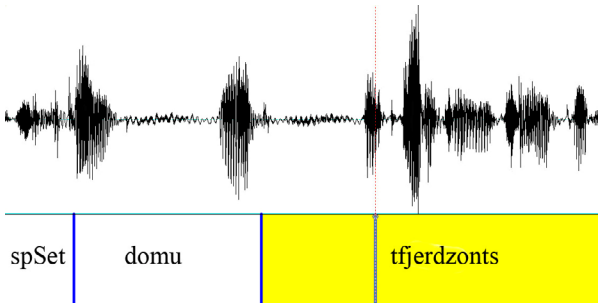

Fig. 8. Text: ‘Sprzed domu twierdząc’ before corrections.

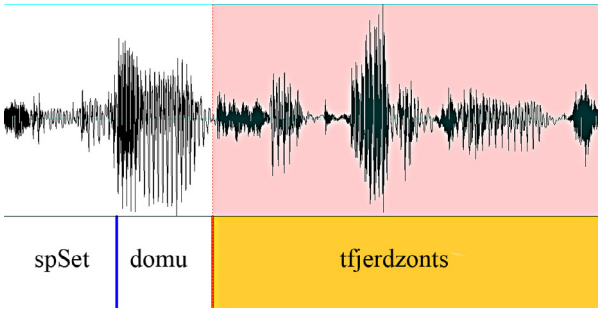

Fig. 9. Text: ‘Sprzed domu twierdząc’ after corrections.

Despite the fact that during the process of conducting the tests there was no intonation model and the cost [15] function was unfinished, the quality of the generated speech turned out to be acceptable and the amount of distortions and errors was on a relatively low level. The occurring distortions didn’t influence the intelligibility, and synthesized speech sounds naturally. It was noticed that appearing distortions of synthesized speech was unavoidable because of automatization degree. Also the corpus was recorded by semi-professional speaker. In the commercial unit-selection systems the corpus is recorded by a professional speaker (actors, radio speakers). This has some influence on the overall quality of synthesized speech.

It is worth to notice, that well prepared parameterization of the signal will cause the better quality of synthesized speech than during the tests. Additionally in Festival [16] the join cost will be prepared and optimized, f0 model will be constructed so the amount of irregularities will fall off by joining better fitted acoustic units.

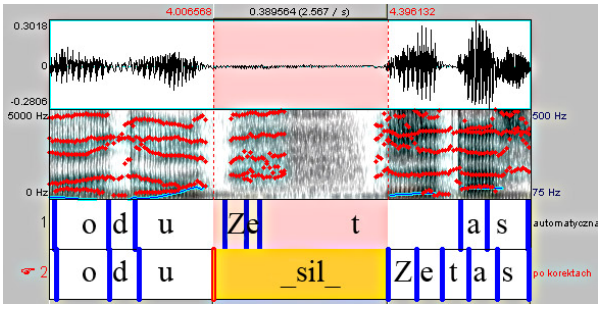

Fig. 10. Comparison of segmentation: automatic and final.

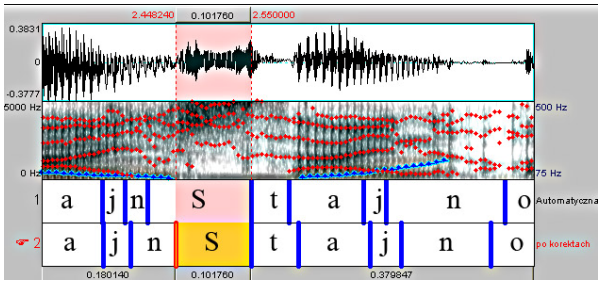

Fig. 11. Other example of manual corrections.

VIII. Conclusion

This paper presents a way for improving the quality of automatic segmentation of acoustic database. All the incorrect boundaries were corrected by moving them to the nearest zero-crossing point. The voiceless plosive phonemes (p, t, k) were automatically corrected and also 4500 phonemes with improper boundaries have been corrected manually. Certainly the problem with a large amount of wrong alignments may be partly caused by existing foreign words in the corpus.

The automatic segmentation turned out to be necessary while working with so many recordings. It was proved that additional automatization was possible and, what is more, it enabled us to eliminate the biggest irregularities and errors in the relatively short time period. It is necessary to notice that the process of the segmentation of acoustic database is very complex. It involves many problems and in order to achieve satisfactory effect it requires a lot of work. However this labour is smaller than in case of the manual segmentation of the large quantity of recordings, which is simply impractical. Despite of bigger accuracy and the smaller amount of "overlooked" errors, manual segmentation of such a large acoustical database is unachievable in a reasonable time. The main advantage of the automatically generated alignment was a significant consequence in setting boundaries, which is practically impossible to obtain during manual segmentation.

It needs to be noted that along with the progress in creating the cost function synthesized speech should sounds more naturally than during the tests. It means that amount of irregularities should be decreased thanks to better selection of acoustic units.

The methods presented in this paper corresponds to the need of large labeled acoustic databases, intended for unit-selection speech synthesis as well as ASR systems.

Acknowledgment

The authors would like to thank prof. Krzysztof Marasek for supplying the grapheme to phoneme converter, Danijel Korzinek and Łukasz Brocki for constructing the prototype synthesis program.

References

[1] K. Szklanny “Preparing the Polish diphone database for speech synthesis in MBROLA”. 50 Otwarte Seminarium z Akustyki Szczyrk, Poland, 2003

[2] R. Clark, K. Simon “Multisyn: Open-domain unit selection for the Festival speech synthesis system”, ScienceDirect, 2007

[3] D. Oliver, K. Szklanny “Creation and analysis of a Polish speech database for use in unit selection synthesis”, LREC Genoa, Italy 2006

[4] K. Marasek, R. Gubrynowicz 2004.”Multi-level Annotation in SpeeCon Polish Speech Database”. In *IMTCI*, pages 58–67.

[5] D. Hirst 1999. “The symbolic coding of segmental duration and tonal alignment: an extension to the INTSINT system”. In *In Proceedings of EUROSPEECH'99*, pages 1639–1642.

[6] A. Black Schultz T. “Speaker Clustering for Multilingual Synthesis” 2006

[7] P. Boersma 2001. “Praat, a system for doing phonetics by computer”. *Glot International*, 5(9/10):341–345.

[8] J. Kominek, A. Black 2004. “Impact of durational outlier removal from unit selection catalogs”. In *Fifth ISCA ITRW on Speech Synthesis (SSW5)*, Pittsburgh, PA, USA.

[9] K. Szklanny, D. Oliver. 2005. “Corpus creation for polish unit selection speech synthesis”. In *InProceedings of Speech Analysis, Synthesis and Recognition: Applications of Phonetics, SASR 2005*, Cracow.

[10] D. Oliver “Polish Text to Speech synthesis system”, MSc Thesis, Edinburgh University, Edinburgh, 1998

[11] K. Richmond, V. Strom, R. Clark, J. Yamagishi, S. Fitt “Festival Multisyn Voices for the 2007 Blizzard Challenge”, Blizzard 2007

[12] M. Kaszczuk, L. Osowski “The IVO software Blizzard 2007 Entry: Improving Ivona Speech Synthesis System”

[13] E. Klabbers, J. P. H. van Santen. 2004. “Clustering of foot-based pitch contours in expressive speech”. In *Fifth ISCA ITRW on Speech Synthesis (SSW5)*, Pittsburgh, PA, USA.

[14] L. Villasenor-Pineda, M. Montes y Gómez, M. A. Pérez- Coutino, and D. Vaufreydaz. 2003. “A Corpus Balancing Method for Language Model Construction”. In *Computational*

[15] N. Campbell “Conversational Speech Synthesis and the Need for Some Laughter”

[16] A. Black, P. Taylor 1998. “Festival Speech Synthesis System: system documentation”. Technical Report HCRC/TR–83, University of Edinburgh, Human Communication Research Centre.